Exploring the Richness and Limitations of Web Sources for Comparable Corpus Research

Gregory Grefenstette
Inria Saclay/TAO, Rue Noetzlin - Bât 660
91190 Gif sur Yvette, France
gregory.grefenstette@inria.fr

Abstract

Comparable Corpora have been used to improve statistical machine translation, for augmenting linked open data, for finding terminology equivalents, and to create other linguistic resources for natural language processing and language learning applications. Recently, continuous vector space models, creating and exploiting word embeddings, have been gaining in popularity in more powerful solutions to creating, and sometimes replacing, these resources. Both classical comparable corpora solutions and vector space models require the presence of a large quantity of multilingual content. In this talk, we will discuss the breadth of this content on the internet to provide some type of intuition in how successful comparable corpus approaches will be in achieving its goals of providing multilingual and cross lingual resources. We examine current estimates of language presence and growth on the web, and of the availability of the type of resources needed to continue and extend comparable corpus research.

Keywords: web mining, under-resourced languages, comparable corpora, language resources

Bibliographical References
